Análise de resíduos nos MLG

Prof. Caio Azevedo

(grande parte do material apresentado foi extraído do livro Modelos de regressão com apoio computacional do Prof. Gilberto A. Paula)

http : //*www.ime.usp.br*/∼*giapaula*/*texto*_2013.*pdf*

Motivação

- Podemos perceber que o resíduo studentizado, muito utilizado para verificar a qualidade de ajuste da classe de MRNLH, dificilmente apresentará normalidade assintótica (embora poderia ser usado para verificar a presença de outliers ou problemas na predição dos valores) para os MLG.
- Portanto, a utilização de outro resíduo se faz necessária.
- Paula (2013) apresenta uma revisão muito boa sobre vários resíduos.
- Nos concentraremos no resíduo componente do desvio (RCD).
- Um fato interessante é que, na definição dos MLG, não aparece nenhum tipo de "erro", como ocorre nos MRNLH.

Introdução

- Um resíduo deve apresentar um comportamento específico quando o modelo está bem ajustado e outro quando o modelo não o estiver.
- O ideal é que, dependendo de qual suposição (ou suposições, p.e., distribuição da variável resposta, independência, função de ligação e forma do preditor linear) não esteja(m) sendo satisfeita(s), alguma mudança específica ocorra em seu comportamento (conforme discutido anteriormente).
- Naturalmente, outras metodologias, para além dos resíduos, podem ser utilizadas para verificar o afastamento de suposições específicas.

RCD

- A forma geral do RCD, para a i-ésima observação, é dada por:

$$T_{D_i} = \frac{d^*(Y_i, \widehat{\mu}_i)}{\sqrt{1 - \widehat{h}_{ii}}} = \frac{\phi^{1/2} d(Y_i, \widehat{\mu}_i)}{\sqrt{1 - \widehat{h}_{ii}}}$$

em que

  - $d(Y_i, \widehat{\mu}_i) = \text{sinal}\,(Y_i - \widehat{\mu}_i)\sqrt{2}\sqrt{D(Y_i; \widehat{\mu}_i)}$.
  - $D(Y_i; \widehat{\mu}_i) = Y_i\left(\widehat{\theta}_i^{(0)} - \widehat{\theta}_i\right) + b(\widehat{\theta}_i) - b(\widehat{\theta}_i^{(0)})$ (em que $\widehat{\theta}_i^{(0)}$ representa o emv sob o modelo saturado e $\widehat{\theta}_i$ o respectivo emv sob o modelo de regressão, ver aula sobre o Desvio).
  - $\widehat{h}_{ii}$ é o i-ésimo elemento da diagonal principal da matriz $\widehat{\boldsymbol{H}} = \widehat{\boldsymbol{W}}^{1/2}\boldsymbol{X}\left(\boldsymbol{X}'\widehat{\boldsymbol{W}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\widehat{\boldsymbol{W}}^{1/2}$, em que $\boldsymbol{X}$ e $\widehat{\boldsymbol{W}}$ são como definidas na parte de estimação.

- Williams (1984) verificou através de simulações que a distribuição de t_{D_i} tende a estar mais próxima da normalidade do que as distribuições de outros resíduos (veja Paula (2013)).
- Utilizando resultados de Cox and Snell (1968), pode-se demonstrar que $\mathcal{E}(D^*(Y_i, \mu_i)) \approx 0$ e $\mathcal{V}(D^*(Y_i, \mu_i)) \approx 1 - h_{ii}$ em que os termos negligenciados são $O(n^{-1})$. Esses resultados reforçam a padronização do RDC por $\sqrt{1 - \widehat{h}_{ii}}$.
- Na prática substituimos ϕ por um estimador consistente (emv, por exemplo).
- A estimativa do RCD é obtida substituindo-se os estimadores nele presentes por suas respectivas estimativas, bem como Y_i pelos valores observados y_i.

Exemplos

- Normal:

$$T_{D_i} = \frac{Y_i - \widehat{\mu}_i}{\sqrt{\widehat{\sigma}^2(1-\widehat{h}_{ii})}}$$

- gama:

$$T_{D_i} = \text{sinal}\,(Y_i - \widehat{\mu}_i)\,\frac{\sqrt{2\widehat{\phi}}}{\sqrt{1-\widehat{h}_{ii}}}\left[-\ln\left(\frac{Y_i}{\widehat{\mu}_i}\right) + \left(\frac{Y_i - \widehat{\mu}_i}{\widehat{\mu}_i}\right)\right]^{1/2}$$

- Bernoulli:

$$T_{D_i} = -\frac{2|\ln(1-\widehat{\mu}_i)|^{1/2}}{\sqrt{1-\widehat{h}_{ii}}}\mathbb{1}_{\{0\}}(Y_i) + \frac{2|\ln(\widehat{\mu}_i)|^{1/2}}{\sqrt{1-\widehat{h}_{ii}}}\mathbb{1}_{\{1\}}(Y_i)$$

- binomial:

$$
\begin{aligned}
T_{D_i} &= \text{sinal}\,(Y_i - \widehat{\mu}_i)\,\frac{\sqrt{2}}{\sqrt{1-\widehat{h}_{ii}}}\Big\{ Y_i \ln[Y_i/(m_i\widehat{\mu}_i)] \\
&+ (m_i - Y_i)\ln\left[(1 - Y_i/m_i)/(1-\widehat{\mu}_i)\right] \times \mathbb{1}_{\{1,\ldots,(m_i-1)\}}(Y_i) \\
&- [m_i|\ln(1-\widehat{\mu}_i)|]\mathbb{1}_{\{0\}}(Y_i) - [m_i|\ln\widehat{\mu}_i|]\mathbb{1}_{\{m_i\}}(Y_i)\Big\}^{1/2}.
\end{aligned}
$$

- Poisson:

$$
\begin{aligned}
T_{D_i} &= \text{sinal}\,(Y_i - \widehat{\mu}_i)\,\frac{\sqrt{2}}{\sqrt{1-\widehat{h}_{ii}}}\left\{Y_i \ln(Y_i/\widehat{\mu}_i) - (Y_i - \widehat{\mu}_i)\right\}^{1/2} I_{\{1,2,\ldots.\}}(Y_i) \\
&\quad \text{sinal}\,(Y_i - \widehat{\mu}_i)\,\frac{\sqrt{2\widehat{\mu}_i}}{\sqrt{1-\widehat{h}_{ii}}} I_{\{0\}}(Y_i).
\end{aligned}
$$

Comentários sobre o RCD

- Pode acontecer de que o modelo esteja bem ajustado e, mesmo assim, a distribuição do RCD pode não ser aproximadamente normal.
- Ainda assim podemos construir uma gráfico de quantil quantil com envelopes simulando a partir do modelo de interesse ao invés da distribuição normal.

Procedimento para se gerar o gráfico de envelopes com o RCD

1) Ajuste o modelo de regressão (estima-se os parâmetros do modelo) obtendo-se as estimativas de MV $(\widetilde{\boldsymbol{\beta}}, \widetilde{\phi})$ e calcule o RCD para cada observação, (t_{D_i}), $i = 1, 2, ..., n$.
2) De posse das estimativas de MV, repita os passos (a) e (b) m vezes.
   a) Simule n variáveis aleatórias ind. $FE(\widetilde{\theta}_i, \widetilde{\phi})$, com $\widetilde{\theta}_i = h(g^{-1}(\widetilde{\boldsymbol{\eta}}_i))$, $\widetilde{\boldsymbol{\eta}}_i = \boldsymbol{X}_i'\widetilde{\boldsymbol{\beta}}$.
   b) Ajuste o modelo de regressão considerando as variáveis simuladas no item a) e obtenha o RCD para cada observação (i) em cada réplica (j).

Procedimento para se gerar o gráfico de envelopes com o RCD

3) Ao final teremos uma matriz com os RCD's, ou seja $t^*_{D_{ij}}$, i=1,...,n, (tamanho da amostra) j=1,...,m (réplica).

$$\boldsymbol{T}_1 = \begin{bmatrix} t^*_{D_{11}} & t^*_{D_{12}} & \dots & t^*_{D_{1m}} \\ t^*_{D_{21}} & t^*_{D_{22}} & \dots & t^*_{D_{2m}} \\ \vdots & \vdots & \ddots & \vdots \\ t^*_{D_{n1}} & t^*_{D_{n2}} & \dots & t^*_{D_{nm}} \end{bmatrix}$$

Procedimento para se gerar o gráfico de envelopes com o RCD

4) Dentro de cada amostra, ordena-se, de modo crescente, os RCD's, obtendo-se $t^*_{D_{(i)j}}$ (estatísticas de ordem):

$$\boldsymbol{T}_2 = \begin{bmatrix} t^*_{D_{(1)1}} & t^*_{D_{(1)2}} & \dots & t^*_{D_{(1)m}} \\ t^*_{D_{(2)1}} & t^*_{D_{(2)2}} & \dots & t^*_{D_{(2)m}} \\ \vdots & \vdots & \ddots & \vdots \\ t^*_{D_{(n)1}} & t^*_{D_{(n)2}} & \dots & t^*_{D_{(n)m}} \end{bmatrix}$$

5) Obtem-se também os limites $t^*_{(i)I} = \min\limits_{1\leq j\leq m} t^*_{D_{(i)j}}$ e $t^*_{(i)S} = \max\limits_{1\leq j\leq m} t^*_{D_{(i)j}}$, $j = 1, 2, ..., m$.

Procedimento para se gerar o gráfico de envelopes com o RCD

5) Na prática considera-se $t^*_{(i)I} = \dfrac{t^*_{D_{(i)(2)}} + t^*_{D_{(i)(3)}}}{2}$ e $t^*_{(i)S} = \dfrac{t^*_{D_{(i)(m-2)}} + t^*_{D_{i(m-1)}}}{2}$ (refinamento das estimativas das bandas de confiança), em que $t^*_{D_{(i)(r)}}$ é a r-ésima estatística de ordem dentro de cada linha, $i = 1, 2, ..., n$.

- Além disso, consideramos como a linha de referência $t^*_{(i)} = \frac{1}{m}\sum_{j=1}^{m} t^*_{D_{(i)j}}, i = 1, 2, ..., n.$

Outros gráficos de interesse

- boxplot/histograma.
- $t_{D_i} \times$ ordem da observação: pontos aberrantes, heterogeneidade (heterocedasticidade) não capturada pelo modelo.
- $t_{D_i} \times g^{-1}(\widetilde{\eta}_i)$(valor predito): pontos aberrantes.
- $\widetilde{z}_i \times \widetilde{\eta}_i$: adequabilidade da função de ligação e do preditor linear (η_i), em que $\widetilde{z}_i = \widetilde{\eta}_i + \widetilde{W}_i^{-1/2}\widetilde{V}_i^{-1/2}(y_i - \widetilde{\mu}_i)$, em que $\widetilde{(.)}$ representa uma estimativa.
- $\widehat{h}_{ii} \times \widehat{\mu}_i$ (pontos alavanca - aqueles que tem um peso desproporcional no próprio valor ajustado, devido à ter um perfil, em termos das covariáveis, diferente dos demais).